H. Metin Aktulga ***Statement of Research Interests***

My research interests are in the area of parallel computing, with emphases on high performance computing, data-intensive computing, and scientific computing. Rapidly growing importance of computational and data-centric methods in science, engineering and industry, as well as the drastic transformations in computer architecture open up exciting opportunities for research in parallel computing. My long-term research goal is to create methods, algorithms and software systems that can harness the full potential of state-of-the-art computing platforms to enable high performance large scale data analytics.

An exciting trend in various fields of science, medicine, business and information technology is the availability of high volume, high velocity and high variety of data, i.e. **big data**. The analyses of large datasets present opportunities for new breakthroughs in these fields. In science, data-centric methods can complement theory, experiment and traditional computational methods or even open up new frontiers where the three pillars of scientific discovery come short of providing satisfactory answers. For example, in the experiments conducted at the Large Hadron Collider to confirm the existence of the Higgs Boson, analysis of extremely large-scale datasets have been crucial. Another example is the detection of supernova explosions which provide important clues for our understanding of the universe. For this purpose, comparative analysis of a vast amount of sky survey data collected over a long time period is necessary. A final example related my own research is the parameterized models used for in-silica study of physical phenomena. Such models are derived through a parameter fitting process against a reference dataset. More accurate models and better predictions can be made possible through the use of large datasets obtained by high throughput screening methods. In any field, collaborative efforts by interdisciplinary research teams are/will be essential to address challenging data-oriented research problems to ensure sustained innovation.

The computational cost associated with the processing of large scale datasets and the need for real-time analytics require the use of high end computing systems, as well as efficient and scalable execution of computational tasks in parallel. As a response to the ever-increasing needs of large scale applications, there has been major transformations in computer architecture recently. On the processor side, multi-core and many-core CPUs have emerged, and GPUs have been introduced for general purpose computing. On the memory side, new interesting technologies such as stack memories and non-volatile memories are entering the marketplace. New network technologies are being developed to ensure fast and reliable communications on massively parallel architectures. In addition to the opportunities in data-driven scientific exploration, the transformations in computer architecture present exciting opportunities in the field of parallel computing.

I will have *two thrust areas* to enable high performance large scale data analytics on emerging architectures. The complexity of the technological stack has grown to an extent where it is imperative to provide frameworks to abstract away the system architecture and orchestration of components for increased productivity. So *my first thrust area* will be to create a software ecosystem that will enable high productivity while still ensuring good performance and scalability for large scale data analysis. The widening gap between computational capabilities and data movement rates of large computer systems, as well as the low arithmetic intensity of typical analytics methods emphasizes parallel algorithms that reduce data movement, utilize asynchrony, amortize communication costs; often at the expense of suboptimal computational complexities and costs. In this respect, *my second thrust area* will be to develop novel parallel algorithms and numerical methods that are both architecture and application-aware to enable efficient and scalable data analytics on emerging architectures.

A key aspect of my research will be close collaborations with domain experts to identify important research problems and to work on solutions together by bringing different perspectives and expertise to the projects. In fact, close interdisciplinary collaboration with various application communities has been a distinguishing aspect of my research during my PhD studies, as well as my subsequent post-doctoral training. I have taken part in major collaborative projects in fields as diverse as molecular modeling, nuclear physics and computational biology. Given the increasing importance of data-driven discovery, I aim to extend these collaborations to colleagues working on machine learning and data sciences. Below I outline my major accomplishments and main future research goals.

Major Accomplishments, Future Research Interests

Software Systems for High Performance and Productive Data Analytics: Petascale scientific computing, next-generation telescopes, high-throughput experiments, and the Internet have been driving a rapid growth in data acquisition and generation. Analyses of these datasets are likely to bring new breakthroughs. Existing solutions such as the widely used Map-Reduce paradigm fail short of providing efficient and scalable analytics for the ever-increasing volume and rate of massive datasets that we are faced with today. The main performance bottleneck of the Map-Reduce framework is in its approach which is completely agnostic to the underlying computer architecture, as well as to the unique characteristics of different applications. On the other end of the spectrum would lie completely custom solutions for each different problem and architecture to achieve the best performance. However, such an approach would stifle productivity and incur very expensive development costs.

Matrix and graph computations are widely used in the analysis of large-scale datasets. I envision a software ecosystem that will enable high productivity while still ensuring good performance by narrowing my focus to matrix and graph based analytics problems. The key ingredient of this ecosystem will be a runtime environment which is aware of both the underlying architecture and the needs of its target applications to determine and execute optimal data movement policies. Data movement in this context refers to I/O from storage systems, moving data between different levels of a deep memory hierarchy and inter-process communication. Most data analysts or scientists are not parallel programming experts, therefore the second key ingredient of the ecosystem that I envision will be an easy-to-use high-level programming interface. By delegating the chore of data movement to a runtime system, the programmer will be able to focus on the computational operations to apply to the subsets of data, and not worry about I/O or communications. This way, I aim to enable an effective interaction between the programmer and the underlying data analytics framework without sacrificing on expressivity, customizability or performance. The final key ingredient will be the artifacts of my research on novel parallel algorithms and numerical methods for emerging architectures which I describe in more detail in the next subsection. They will serve as high performance computational kernels in the form of a data analytics library.

The task-based data-flow middleware, DOoC+LAF, that we developed for large-scale iterative solvers on SSD-equipped clusters [ICCPW12, Cluster12] is an excellent preliminary example of the ecosystem that I describe here. In close collaboration with scientists working on data-driven methods for scientific discovery, I aim to develop our preliminary work into the full fledged data analytics ecosystem that I described above.

Finally, what are the performance bottlenecks in existing and proposed computer architectures for data-intensive applications and how can those bottlenecks be mediated? In the spirit of co-design in collaboration with computer architecture experts, I aim to provide feedback from the perspective of real-world applications to the design of next generation computer systems for large-scale data analytics. Our work which received a best paper nomination in the prestigious SC13 conference is a prime example of such an effort.

Parallel Algorithms and Numerical Methods for Emerging Architectures: While the computational power of high-end computer systems has been increasing steadily for decades, the capacity and bandwidth of their memory subsystems which is critical for the performance of large scale data-intensive applications have not been able to keep pace. As we move forward, this gap is anticipated to widen even further. The main reason for this trend is that it is not possible to meet the storage capacity and power consumption requirements of large computer systems using the existing DRAM technology. As a result, the memory hierarchy will get deeper with stack memory and non-volatile memory (NVM) solutions that feature higher storage densities and lower power requirements being key components of future architectures. We are already seeing this change in the form of high performance flash memory storages on modern clusters. Equally important are the changes in the processor architecture. Future processors are expected to host hundreds of heterogenous cores which may not have full cache-coherency anymore. Current parallel algorithms and numerical methods have not been designed and implemented for these drastically different architectures. New algorithmic solutions are needed to leverage the full potential of the technological advancements in computer architecture.

My previous research accomplishments include several examples of architecture-aware algorithm de-

sign. For example, in [EuroPar12,CPE13], I have designed a highly scalable implementation of the well-known Lanczos algorithm to find extremal eigenpairs of very large sparse matrices on multi-core architectures. I have used a topology-aware mapping of processes to physical processors to reduce the load on the network. Communication overheads were reduced even further by exploiting the multi-core processor architecture to overlap expensive collective communication operations with computations at the expense of a slight increase in computational complexity. As part of my ongoing work, which is partially reported in [IPDPS14-1], I am investigating the use of block iterative methods in place of the more commonly used Krylov subspace methods for iterative eigenvalue computations. Even though block iterative methods typically converge slower, they expose more parallelism and feature better data locality in the performance critical sparse matrix computations part, so they can harness much better performance out of modern CPUs. In the context of nuclear configuration interaction computations, we have shown that block iterative methods can be up to 5x faster in the sparse matrix vector multiplication computations, providing inspiration to use a block eigensolver instead of the Lanczos algorithm. Another example is our work on the development of a novel eigensolver library based on the spectral partitioning approach, named *Eigenbuster* [ParCo13]. In our approach, a coarse grained level of parallelism is introduced by partitioning the spectrum of interest into sub-intervals. This method increases the total amount of concurrency in computation and maps well to the massively parallel supercomputers available today. Initial performance evaluations suggest that the approach that we adopt in *Eigenbuster* is a promising one to compute a large number of eigenpairs of big sparse matrices compared to the existing methods that try to extract all eigenpairs from a single subspace.

In my research, I will work on designing novel parallel algorithms and numerical methods that pay attention to the deep memory hierarchies and massive on-chip parallelism of the emerging computer architectures. My main focus area will be on matrix and graph computations, because they are so fundamental to several problems in large-scale data analytics. Power is a big concern for future computer systems with data movement being a major source of energy consumption. So for future systems with deep memory hierarchies, can we design algorithms that increase data locality and reduce data movement, or even recompute some set of data instead of moving it back and forth between various layers of the memory system? NVM realizations exhibit very different characteristics than existing technologies because their performances and energy consumptions during write operations can be much worse compared to read operations. Are there *write-reducing* algorithms for the numerical methods commonly used in data analytics? What are the best data structures to store and access massive datasets on stack memories and NVM devices which, unlike the DRAM memory, feature a high degree of parallelism within the device itself? Message passing based parallelism incurs expensive communication overheads, while most existing thread-level parallel programming models assume cache-coherency. Can we instead implement our algorithms and numerical methods more efficiently on a massively parallel chip without full cache-coherency using the emerging PGAS languages? These are some sample research questions that I will work to find answers and provide solutions.

Data-driven Numerical Optimization Techniques for Molecular Modeling & Simulation: Molecular modeling is a powerful tool for simulating and understanding diverse systems – ranging from materials processes to biophysical phenomena. As part of my PhD studies, I have developed a highly scalable parallel reactive molecular dynamics code, PuReMD [SISC12,ParCo12], which is now actively being used by hundreds of researchers worldwide. During my postdoc, I have been worked on eigensolvers for fast and scalable electronic structure computations [ParCo13].

While electronic structure computations provide highly accurate results, their applicability is limited in terms of system sizes and simulation time-frames. Classical molecular dynamics (MD) methods, on the other hand, allow the study of systems of much larger scales by significantly reducing the number of degrees of freedom in the problem. The quality of the results obtained from classical MD simulations, to a great extent, depends on the quality of the functions describing the interactions and their parametrizations. A high quality classical MD force field can have a huge impact in the design and simulation of advanced materials. However, generation and tuning of force fields is a very labor-intensive task, requiring deep expert knowledge and years of dedicated efforts. Some computational scientists spend their entire lives on the generation and optimization of force fields to study certain systems.

I envision a fully automated force field generation and optimization framework to streamline the design and simulation of advanced materials. I aim to lessen or even completely remove the dependence on human experts by leveraging the rapid proliferation of experimental and computational data that are becoming available under the Materials Genome Initiative and the vast computational power of modern supercomputers. My research on data-driven optimization techniques will be supplemented by my research described in the previous subsections: parallel algorithms and numerical methods that I develop for emerging architectures will enable efficient high-throughput screenings of materials, as well as fast numerical optimization techniques; the high performance data analytics framework will be crucial for the management and analysis of large-scale experimental and computational datasets.

Two main tasks are involved in force-field optimization. First one is the identification of a reference dataset which can accurately characterize the molecular system of interest. The second one is the optimization of the functional forms and parameter sets in order to approximate the reference dataset as best as possible. Current methods can be described more like an art rather than a systematic approach due to the difficulty of identifying a good reference dataset and the hardness of the optimization problem involved. There exists efforts for automating the force-field optimization process, but these efforts largely remain specific to some molecular systems of interest, and force fields that are being used to study them.

My approach will be agnostic to the force-field and the molecular system of interest. Towards this end, it will draw ideas and techniques from several fields of computer science and applied math. Machine learning and data mining techniques will be used to identify a reference dataset from two major sources: (i) information accumulated through previous experimental as well as computational studies, (ii) high-throughput screenings based on density-functional theory (DFT) simulations. Due to the large volume and variety of scientific data involved, efficient and scalable data management and analysis techniques will be key research topics. Sensitivity analysis and uncertainty quantification techniques are essential to determine the set of parameters that need to be tuned to obtain the best agreement with the reference dataset. Further along the road, I aim to investigate the question of whether it is possible to automatically generate a set of force field functions to come up with high quality descriptions of the molecular systems of interest, rather than relying on the ones provided by the computational scientists.

Towards this end, I have initiated collaborations with the Materials Project and the Electrolyte Genome groups at LBL, which are large DOE funded projects under the Materials Genome Initiative. I aim to continue these collaborations as a faculty member and seek for large interdisciplinary research funding from DOE and DARPA.

Summary

Large-scale data analytics hold a great promise for addressing the most challenging problems that our society faces today. In my research, I aim to create methods, algorithms and software systems that can harness the full potential of large scale computer systems to address challenging problems in this area. In this respect, major transformations in computer architecture and the ever-increasing volume, rate and variety of data in various fields present exciting research opportunities in parallel computing. Interdisciplinary collaborations with experts from various fields will be an essential part of my research agenda. The research experience I had during my PhD and postdoctoral training years have provided me the background, skill sets and collaborations to achieve my long-term research goals.

[CLUSTER12] Z. Zhou *et al.* "An out-of-core eigensolver on SSD-equipped clusters," in *Proc. Cluster 2012*.
[ICPPW12] —-, "An out-of-core dataflow middleware to reduce the cost of large scale iterative solvers," in *Proc. ICPPW 2012*.
[SC13] M. Jung *et al.* "Exploring the future of out-of-core computing with compute-local non-volatile memory", in *Proc. SC'13*.
[EuroPar12] H. M. Aktulga *et al.* "Topology-aware mappings for large-scale eigenvalue problems," in *Proc. Euro-Par 2012*.
[CPE13] —-, "Improving the scalability of a symmetric iterative eigensolver for multi-core platforms," *Concurrency and Computation: Practice and Experience*, published online, Sep 2013.
[IPDPS14-1] H. M. Aktulga *et al.* "Performance optimization of block eigensolvers for CI calculations," *submitted to IPDPS 2014*.
[ParCo13] H. M. Aktulga *et al.* "Parallel eigenvalue calculation based on multiple shift-invert Lanczos and contour integral based spectral projection method," *submitted to Parellel Computing*.
[ParCo12] H. M. Aktulga *et al.* "Parallel reactive MD: Numerical methods and algorithmic techniques," *Parallel Comp.*, 2012.
[SISC12] —-, "Reactive MD: Numerical methods and algorithmic techniques," *SIAM Journal on Sci. Comp.*, 2012.